Stylometric Analysis of Parliamentary Speeches: Gender Dimension

Justina Mandravickaitė
Vilnius University, Lithuania
Baltic Institute of Advanced
Technology, Lithuania
justina@bpti.lt

Tomas Krilavičius
Vytautas Magnus University, Lithuania
Baltic Inistitute of Advanced
Technology, Lithuania
t.krilavicius@bpti.lt

Abstract

Relation between gender and language has been studied by many authors, however, there is still some uncertainty left regarding gender influence on language usage in the professional environment. Often, the studied data sets are too small or texts of individual authors are too short in order to capture differences of language usage wrt gender successfully. This study draws from a larger corpus of speeches transcripts of the Lithuanian Parliament (1990–2013) to explore language differences of political debates by gender via stylometric analysis. Experimental set up consists of stylistic features that indicate lexical style and do not require external linguistic tools, namely the most frequent words, in combination with unsupervised machine learning algorithms. Results show that gender differences in the language use remain in professional environment not only in usage of function words, preferred linguistic constructions, but in the presented topics as well.

1 Introduction

Gender influence on language usage have been extensively studied (Lakoff, 1973; Holmes, 2006; Holmes, 2013; Argamon et al., 2003) without fully reaching a common agreement. Understanding gender differences in professional environment would assist in a more balanced atmosphere (Herring and Paolillo, 2006; Mullany, 2007), however results on extent of variation depending on context of communication in professional setting are inconclusive(Newman et al., 2008).

Most studies rely on the relatively small data sets, or texts of the individual authors are too short to capture the differences in the language due to the gender (Newman et al., 2008; Herring and Martinson, 2004). Some results show that gender differences in language depend on the context, e.g., people assume *male language* in a formal setting and *female* in an informal environment (Pennebaker, 2011). We investigate gender impact to the language use in a professional setting, i.e., transcripts of speeches of the Lithuanian Parliament debates. We study language wrt style, i.e., *male* and *female* style of the language usage by applying computational stylistics or stylometry. Stylometry is based on the two hypotheses: (1) *human stylome hypothesis*, i.e., each individual has a unique style (Van Halteren et al., 2005); (2) unique style of individual can be measured (Stamatatos, 2009), stylometry allows gaining meta-knowledge (Daelemans, 2013), i.e., what can be learned from the text about the author - gender (Luyckx et al., 2006; Argamon et al., 2003; Cheng et al., 2011; Koppel et al., 2002), age (Dahllöf, 2012), psychological characteristics (Luyckx and Daelemans, 2008), political affiliation (Dahllöf, 2012), etc.

Like in most studies of gender and language (Yu, 2014; Herring and Martinson, 2004), biological sex as a criterion for gender was used in this study. We compare differences of the gender related language use at the group level (faction). Lithuanian language allows easy distinction between male and female legislators based on their names in the transcripts.[1]

We investigate several questions: (1) How well simple stylistic features distinguish genders of members the Lithuanian Parliament? (2) Which differences in language use by female and male Lithuanian Parliament members selected features and methods are able to capture?

[1] Of course, all information about members of parliament is available on-line.

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 102–107,
Valencia, Spain, 4 April 2017. ©2017 Association for Computational Linguistics

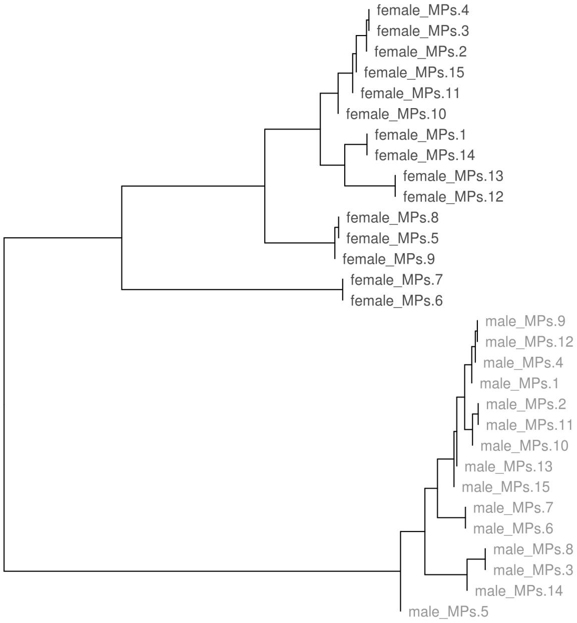

Figure 1: Results with 7000 MFW as features.

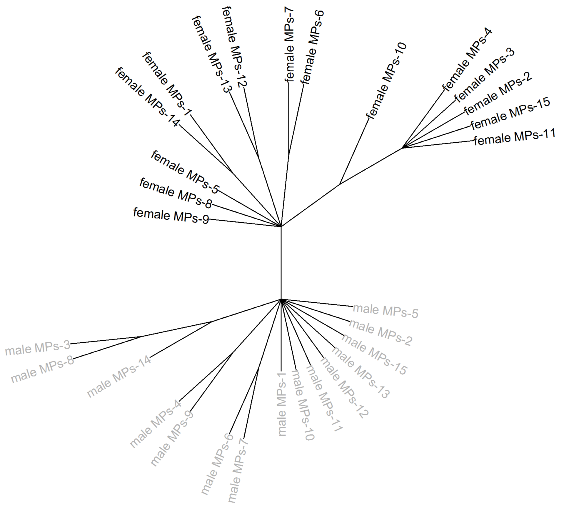

Figure 2: Bootstrap Consensus Tree with Canberra and 100–10000 MFW.

2 Data Set

Corpus of parliamentary speeches in the Lithuanian Parliament[2] is used. It consists of transcripts of parliamentary speeches from March 1990 to December 2013, 10727 of female members of Parliament (MPs) and 100181 of male MPs, overall 23 908 302 words (2 357 596 of female MPs and 21 550 706 of male; see Table 2 for the details). Only speeches of at least 100 words and of MPs with at least 200 of them were included in the corpus (Kapočiūtė-Dzikienė and Utka, 2014). It could have diminished number of female MPs speeches included into the corpus and our analysis as well. However, the choice of unsupervised learning approach downscales class imbalance problem, i.e. significant difference in number of transcribed parliamentary speeches made by female and male MPs.

Lithuanian is a highly inflective language, i.e. nouns have grammatical gender, number and semantic relations between them are expressed with 7 cases; adjectives have to match nouns in terms of gender, number and case; verbs have 4 tenses and particles for each of them, with ending marking its tense, person and number; gender and case for the particles are also marked morphologically at the ending. All these features produce a substantial number of inflective forms for one lemma. Thus in order to avoid data sparseness we did not lemmatize corpus for our experiments.

To get around of "fingerprint" of individual authorship as much as possible, all the samples were concatenated into two large documents based on the gender, and then were partitioned into 15 parts each. Thus for analysis we had 15 samples of parliamentary speech made by female MPs and another 15 samples – made by male MPs.

3 Stylistic Features and Statistical Measures

We use the most frequent words (MFW) (Burrows, 1992; Hoover, 2007; Eder, 2013b; Rybicki and Eder, 2011; Eder and Rybicki, 2013; Eder, 2013a) (usually, they coincide with function words (Hochmann et al., 2010; Sigurd et al., 2004)), as features, because they are considered to be topic-neutral and perform well (Juola and Baayen, 2005; Holmes et al., 2001; Burrows, 2002).

Stylo package for stylometric analysis using R (Eder et al., 2014) is used for experiments.

Experiments are performed in batches using different number of MFWs, firstly, using the whole corpus, raw frequency list of features is generated, then normalized using *z-scores*, which measure distance of features frequencies in the corpus in terms of their proximity to the mean (Hoover, 2004), where z-scores are defined as $z = \frac{A_i - \mu}{\sigma}$, where A_i is *frequency* of a feature, μ is *mean fre-*

[2] Corpus of parliamentary speeches in the Lithuanian Parliament was created in the project "Automatic Authorship Attribution and Author Profiling for the Lithuanian Language" (ASTRA) (No. LIT-8-69), 2014 – 2015.

MPs by gender	No. of samples	No. of words	No. of unique words
Female	10 727	2 357 596	93 611
Male	100 181	21 550 706	268 030

Table 1: Statistics of Corpus of parliamentary speeches in the Lithuanian Parliament.

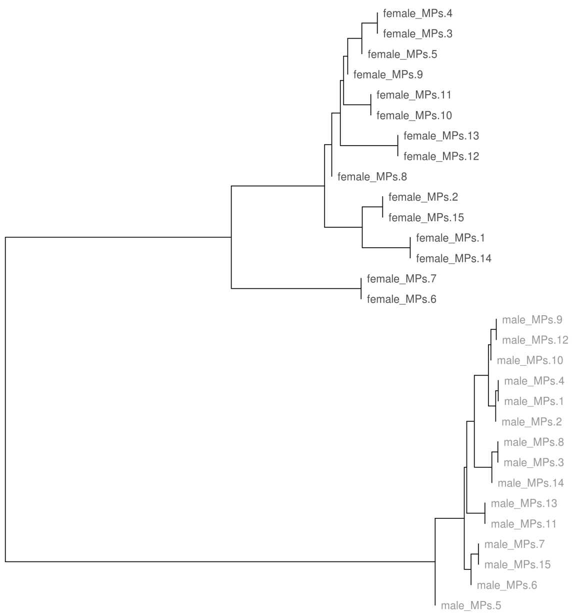

Figure 3: Results with 200 MFW (starting at 6800 MFW).

quency of certain feature in one document, σ is a standard deviation.

Dissimilarity between the text samples is calculated using selected distances (see below), and distance matrix is generated. Then, *hierarchical clustering* is applied to group samples by similarity (Everitt et al., 2011), and dendrograms are used to visualize the results.

Typically Burrows's Delta distance is used for stylometric analysis (Burrows, 2002; Rybicki and Eder, 2011). However, Delta depends on *z-scores*, number of documents and balance of terms in documents, length and number of authors (Stamatatos, 2009). While Burrow's Delta is effective for English and German, it is less successful for highly inflective languages, e.g., Latin and Polish (Rybicki and Eder, 2011). Hence we used Eder's Delta, i.e., a modified Burrows's Delta that gives more weight to the frequent features and rescales less frequent to avoid random infrequent ones (Eder et al., 2014). It was defined to use with highly inflected languages, such as Lithuanian. However, we have achieved the best results with Canberra distance $\delta_{(AB)} = \sum_{i=1}^{n} \frac{|A_i - B_i|}{|A_i|+|B_i|}$ where n is a number of most frequent features, A and B are documents, A_i and B_i are frequencies of a given feature in the documents A and B in the corpus, respectively (Eder et al., 2014). It was reported to be suitable for inflective languages, albeit it is sensitive for rare vocabulary (Eder et al., 2014), e.g., words that occurred only once or twice.

The goal is identifying stylistic dissimilarities and mapping positions of the text samples in relation to each other, not classifying female/male legislators, hence hierarchical clustering with Ward linkage (it minimizes total variance within-cluster (Everitt et al., 2011)) was chosen. Though it is sensitive to changes in a number of features or methods of grouping (Eder, 2013a; Luyckx et al., 2006), in this study it shows stable results. Robustness of clustering results was examined using bootstrap procedure (Eder, 2013a). It includes extensions of Burrows's Delta (Argamon, 2008; Eder et al., 2014) and bootstrap consensus trees (Eder, 2013a) as a way to improve reliability of cluster analysis dendrograms.

4 Experiments

From 20 to 10 000 most frequent features were used for each experiment. We use hierarchical clustering with Ward linkage and Canberra distance, and visualize results in dendrograms to map positions of the samples in relation to each other.

We focus on identifying variation in female and male parliamentary speech, and do not analyze smaller clusters and dynamics inside them. A more detailed investigation of separate features (e.g., specific words, part-of-speech tags or their sequences) that are characteristic to female MPs and male MPs individually, are part of future plans, while in this paper we focus on the most frequent words.

Experiments with more MFW (from 7000 up to 9910) successfully separated samples of parliamentary speeches by gender, see Figure 1. *Bootstrap Consensus Tree* (BCT) procedure (hierarchical clustering and aggregation of results into con-

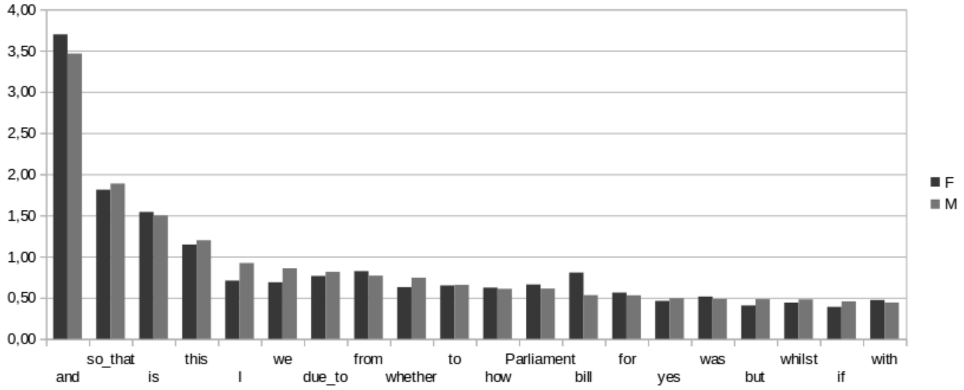

Figure 4: 20 MFW from the beginning with normalized frequencies.

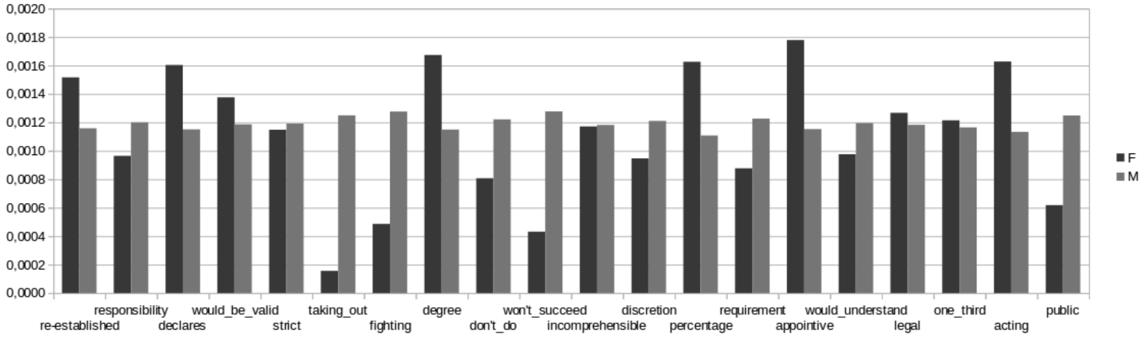

Figure 5: 20 MFW from the range of lesser frequency (6880–7000 MFW).

sensus tree (Eder, 2013a)) was applied to analyze the results. Consensus strength of 0.75 was chosen, i.e., the two documents are related, if they are related in the same proportion in the hierarchical clustering. So, consensus strength 0.75 means that visualized linkages appear in at least 75% of the clusters. See Figure 2 for BCT results for separating male and female legislators in the Lithuanian Parliament.

We needed at least 7000 MFW for clear differentiation of parliamentary speeches by gender in LT parliament. It shows that differences in topics presented as content words are less frequent than function words. To test this assumption, we performed experiments with different number and ranges of MFWs. As Figure 3 shows, less frequent MFWs capture gender variation as well.

The following gender based differences were noted male speeches transcripts (underscores show merge words that are one word in Lithuanian, but are several in English): (1) pronouns *I, we*; (2) demonstratives (e.g. *this*); (3) conjunctions *but, whether, if*; (4) negations (*won't_succeed, don't_do*); (5) *responsibility, public*; (6) *fighting, taking_out*. Some common characteristics of transcripts of female speeches: (1) conjunction *and*; (2) preposition *with*; (3) *parliament, bill*; (4) measurements (*degree, percentage*); (5) parliamentary procedures (*acting, appointive, would_be_valid, legal*). See Figures 4 and 5 for details.

The results show that simple features and methods, such as MFW and hierarchical clustering, perform well with Lithuanian (morphology-rich language with relatively free word order, thus, challenging for many NLP tasks) and identify gender effect on language variation in LT parliament speeches transcripts, and do not require using lemmas (Kapočiūtė-Dzikienė et al., 2014), part-of-speech n-grams (Eder, 2010) and other feature combinations (Argamon et al., 2007; Argamon et al., 2003; Yu, 2014)).

5 Conclusion and Future Work

Results show that MFW and hierarchical clustering with Canberra distance successfully capture variation in transcripts of speeches by female and male MPs, which are clearly visible in dendrograms. Experiments with different ranges of MFW show, that more frequent MFW identify variation in usage of function words, medium frequent MFW reveal variation in topics presented. Thus, for female MPs conjunction *and*, preposition *with*, words *parliament* and *bill*, words for measuring and parliamentary procedures were more characteristic, while male MPs tended to use more first person pronouns, demonstratives, negations, conjunctions *but, whether, if* and words *responsibility, public, taking out, fighting*.

Future plans include experiments with different domain documents, diverse language types (e.g., formal, informal), investigation of other features (e.g., specific words, lemmas, part-of-speech tags or their sequences) that are characteristic to different genders, and other distance measures.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *To appear in Text*, 23:3.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Shlomo Argamon. 2008. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147.

John F. Burrows. 1992. Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7(2):91–109.

John Burrows. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):78–88.

Walter Daelemans. 2013. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, pages 451–462. Springer.

Mats Dahllöf. 2012. Automatic prediction of gender, political affiliation, and age in swedish politicians from the wording of their speeches - a comparative study of classifiability. *Literary and linguistic computing*, 27(2):139–153.

Maciej Eder and Jan Rybicki. 2013. Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 28(2):229–236.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2014. Package 'stylo'.

Maciej Eder. 2010. Does size matter? authorship attribution, small samples, big problem. *Proceedings of Digital Humanities*, pages 132–135.

Maciej Eder. 2013a. Computational stylistics and biblical translation: How reliable can a dendrogram be. *The translator and the computer*, pages 155–170.

Maciej Eder. 2013b. Mind your corpus: systematic errors in authorship attribution. *Literary and linguistic computing*, 28(4):603–614.

Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. Hierarchical clustering. *Cluster Analysis, 5th Edition*, pages 71–110.

Susan C. Herring and Anna Martinson. 2004. Assessing gender authenticity in computer-mediated language use evidence from an identity game. *Journal of Language and Social Psychology*, 23(4):424–446.

Susan C. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Jean-Rémy Hochmann, Ansgar D. Endress, and Jacques Mehler. 2010. Word frequency as a cue for identifying function words in infancy. *Cognition*, 115(3):444–457.

David I. Holmes, Lesley J. Gordon, and Christine Wilson. 2001. A widow and her soldier: Stylometry and the american civil war. *Literary and Linguistic Computing*, 16(4):403–420.

Janet Holmes. 2006. Sharing a laugh: Pragmatic aspects of humor and gender in the workplace. *Journal of Pragmatics*, 38(1):26–50.

Janet Holmes. 2013. *Women, men and politeness*. Routledge.

David L. Hoover. 2004. Delta prime? *Literary and Linguistic Computing*, 19(4):477–495.

David L. Hoover. 2007. Corpus stylistics, stylometry, and the styles of henry james. *Style*, 41(2):174.

Patrick Juola and R. Harald Baayen. 2005. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20(Suppl):59–67.

Jurgita Kapočiūtė-Dzikienė and Andrius Utka. 2014. Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. *Linguistics: Germanic & Romance Studies/Kalbotyra: Romanu ir Germanu Studijos*, 66.

Jurgita Kapočiūtė-Dzikienė, Ligita Sarkute, and Andrius Utka. 2014. Automatic author profiling of Lithuanian parliamentary speeches: Exploring the influence of features and dataset sizes. In *Human Language Technologies - The Baltic Perspective -*

Proceedings of the Sixth International Conference Baltic HLT 2014, Kaunas, Lithuania, September 26-27, 2014, pages 99–106.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(01):45–79.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Kim Luyckx, Walter Daelemans, and Edward Vanhoutte. 2006. Stylogenetics: Clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy*.

Louise Mullany. 2007. *Gendered Discourse in the Professional Workplace*. Communicating in Professions and Organizations. Palgrave Macmillan UK.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

James W. Pennebaker. 2011. The secret life of pronouns. *New Scientist*, 211(2828):42–45.

Jan Rybicki and Maciej Eder. 2011. Deeper delta across genres and languages: do we really need the most frequent words? *Literary and linguistic computing*, 26(3):315–321.

Bengt Sigurd, Mats Eeg-Olofsson, and Joost Van Weijer. 2004. Word length, sentence length and frequency–zipf revisited. *Studia Linguistica*, 58(1):37–52.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Bei Yu. 2014. Language and gender in congressional speech. *Literary and Linguistic Computing*, 29(1):118–132.